Neurotoxicology and Teratology, Vol. 14, pp. 221–228, 1992
Printed in the U.S.A.

0892-0362/92 $5.00 + .00

Principles and Pitfalls in the Analysis of Prenatal Treatment Effects in Multiparous Species

R. R. HOLSON[1] AND B. PEARCE

Division of Reproductive and Developmental Toxicology, National Center for Toxicological Research, Jefferson, AR 72079

Accepted 4 March 1992

HOLSON, R. R. AND B. PEARCE. *Principles and pitfalls in the analysis of prenatal treatment effects in multiparous species.* NEUROTOXICOL TERATOL 14(3), 221–228, 1992.—Developmental studies often assess the effect of treatment of the pregnant mother on offspring. The use of multiparous species such as rats and mice in such studies creates a special set of design and analysis problems. These arise for two reasons. First, the availability of many offspring per litter tempts the experimenter to inflate sample size by treating scores from several pups per litter as independent observations. Second, large litter size seldom makes it practical to measure exposure effects in all offspring of an exposed dam. Such studies commonly involve two-stage sampling: Drawing a random sample of dams for treatment, then drawing a second sample of pups per dam for neurobehavioral measurements. In this article, such sampling was modeled by two different simulations. The first, a standard *Monte-Carlo approach*, sampled from random-normal distributions for litter mean and within-litter variability. The second simulation sampled without replacement from actual data on weight of all pups in a series of 39 nontreated rat litters. These mutually-supportive approaches demonstrate that litter effects, even over as few as three litters, are generally large and statistically meaningful. Consequently, statistical significance tests are sensitive to litter effects. Inflation of sample size by treating as few as 2 pups per litter as independent measurements can almost triple the nominal 0.05 alpha level. Furthermore, two-stage sampling increases the within-treatment error term and correspondingly reduces statistical power relative to one-stage sampling. Calculating litter means from measurements of as few as 2 pups per litter can increase power but never to the degree obtained by holding number of pups/treatment constant, measuring just one pup/litter and correspondingly increasing the number of litters per treatment.

Neurobehavioral teratology Experimental design Independence Statistical assumptions Litter effects Power Violation of assumptions

DIRECT experimental intervention in prenatal development is an important tool in developmental research. In some cases such studies directly treat the embryo or fetus, affecting the mother indirectly if at all. These techniques include virtually all in vitro approaches, direct treatment of the conceptus and procedures such as inducing hypoxia by briefly ligating blood vessels to one horn of the uterus (5). Conversely, many experiments treat the dam directly, and through her the conceptus. Studies of this kind often seek to model drug exposure or insults seen in pregnant humans and are the most common type of study in the nascent area of neurobehavioral teratology. It is this latter set of techniques that we are concerned with in this article.

Humans and most primates give birth to a single infant. In such cases the treated maternal-fetal unit forms an n of 1. In contrast are multiparous species, especially that most common experimental model, the laboratory rodent. Here a substantial number of offspring are produced by a single dam. This fecundity gives rise to more than one problem in experimental design and analysis, problems that are the focus of this article.

To better understand the effect of multiple births on experimental outcome, it is necessary to briefly consider the question of within and between-litter variance. With one infant per pregnancy there can be no within-litter variance. With more than one infant per pregnancy, it is possible to derive a within-litter error term, with which one can measure the magnitude of between-litter variance, referred to simply as *litter effects*. It is generally assumed that pups within a litter are more alike than are pups compared across litters. This is a plausible assumption, because pups within litters are genetically alike and certainly share a common intrauterine environment. To the degree that this is true, litter effects (the difference between litters relative to differences within litters), is

[1] To whom requests for reprints should be addressed.

large. This supposition is borne out by studies that have directly measured such effects (2,8,9).

The (probable high) magnitude of litter effects in multiparous species has direct bearing on the severity of a common analysis flaw in studies of the effects of prenatal treatment of the dam. This is the use of more than 1 score per litter in statistical analyses of treatment effects, when such analyses do not include or correct for litter effects. This practice ignores litter effects by treating each score as independent. This error is referred to here as *inflated sample size*, or simply inflated *n*. Technically, this practice is a violation of the assumption of independence in significance testing and is known to carry the risk of substantial inflation of the nominal 0.05 alpha level (1,8,9,11). More intuitively, one can easily grasp how large litter effects might create spurious false-positives in conjunction with inflated *n*. In an extreme example, assume an *n* of 12 per group in each of two groups, a control and a treated group. Furthermore, assume that all 12 treated pups are from a single treated dam and likewise for the control pups. It is clear that in this example treatment effects are completely confounded with litter effects. Consequently (given substantial litter effects), there is a high probability of concluding that the experimental treatment produced a difference which was due to innate differences between litters. In short, the existence of more than 1 pup per treated maternal-fetal unit carries with it both the opportunity to measure litter effects and the temptation to confuse such effects with treatment effects by inflating sample size.

There are two closely related strategies available to the researcher to avoid spurious inflation of sample size. First, and theoretically most attractive, one can include litter as an independent, random, and nested factor in the analysis. This approach specifically controls for litter effects, while offering the investigator a direct statistical test of the significance of such effects (See Hughes, 1979 (8) for a full and lucid presentation of this approach). However, this approach is only feasible with 2 or more scores per litter. The second alternative uses 1 score per litter, either a litter mean or the score of a single animal per litter. The two approaches are mathematically identical in their test for treatment effects (giving identical *F* ratios and probabilities for treatment effects), but because the use of litter means is conceptually simpler and more general (extending to the limiting case of one pup per litter), this latter approach is presented here.

Another consequence of having a substantial number of progeny per treated dam is that in most studies experimenters cannot practically conduct outcome measures on all offspring. Instead, some subset of pups are assessed in each litter. This practice entails two-stage sampling, i.e., the experimenter first draws a sample of pregnant dams for treatment and then draws a second sample of treated pups from within each litter for neurobehavioral assessment. This two-stage sampling complicates traditional statistical models of variance by introducing expressions for both within- and between-litter variance into the error term (3). The consequence is to increase variance within treatment conditions, and to reduce power below that for the more traditional single-sample statistical models.

All of the aforementioned considerations are addressed in this article. We shall apply a combination of approaches, including two types of simulation: Monte Carlo simulations and simulations from actual data sets obtained by measuring gestational day (GD) 20 weights for all fetuses in 39 untreated rat litters.

CONCEPTS AND DEFINITIONS

Prior to introducing study methodology it will be helpful to introduce terminology and several important concepts.

Terms are defined as follows:

- S = Number of simulations
- C = Number of cells or conditions (e.g., control, low & high dose, C = 3)
- L_c = Number of Litters per Condition
- P_L = Number of pups sampled per Litter
- $n_c = L_c{}^*p_L$ or total scores per condition
- $N = C^*L_c{}^*p_L$ or total scores per experiment
- VAR_L, SD_L = Variance or SD of litter means
- VAR_p, SD_p = Variance or SD of pups within litters, averaged over all litters.
- Z = Effect size, in MS_w or SD units. A Z of 2.0 means that the maximal difference between control mean and treatment mean is 2 SDs or $2^*(MS_w)^{1/2}$

Readers are doubtlessly familiar with the basic concepts of analysis of variance (ANOVA) (10,15). To determine whether treatment groups or litters differ on some measure, an *F* ratio is calculated. This *F* ratio is a measure of variance between groups or litters (mean square between or MS_B), divided by a measure of the average variance within all groups or litters (mean square within or MS_W). Symbolically,

$$F = MS_B/MS_W$$

This deceptively simple equation does hide certain complexities, especially where MS_B is concerned. This measure of variability is not simply the variance of treatment group means or litter means. Rather, it is the variability between group means multiplied by the number of subjects or scores per group. In other words, if we symbolize the variance of group means as VAR_B and the number of scores per group (assuming equal sample size in all groups), then as

$$MS_B = n_c{}^*VAR_B$$

i.e., to calculate the mean square between treatment groups, one assigns the variability between group means to each score or subject.

In terms of litter effects, MS_B has little intuitive appeal. VAR_L, the variance between litter averages, is a more easily-interpreted measure of the magnitude of differences between litters (This is why Cohen (4) uses the ratio SD_B/SD_W as a measure of effect size). In the case of litter effects, if p_L is the number of pups measured per litter, we write

$$MS_B = p_L{}^*VAR_L$$

and for mean square within, the variance within a litter averaged over all litters, we write

$$MS_W = VAR_P$$

What is a likely range of actual biological values of VAR_L (the variability between litter means) relative to VAR_P, the mean variability within litters? Intuitively, in the laboratory setting most multiparous dams are genetically interrelated, in that most laboratory rodents are inbred. Consequently, dams will, at best, be as like or different as are pups within a single litter. In other words, we expect that

$$VAR_L \approx VAR_p \text{ and that } MS_L \approx MS_p$$

This expectation is confirmed by our data on fetal weight in 39 untreated rat litters. Here we find that:

$$MS_L = VAR_L = 0.089$$

$$MS_p = VAR_p = 0.077.$$

Setting $VAR_L \approx VAR_p$ has profound implications for the magnitude of litter effects. In single-stage sampling,

$$MS_w = VAR_p$$

$$MS_B = p_L{}^*MS_L$$

where p_L = the number of pups sampled per litter. Consequently, in testing the significance of litter effects, we obtain:

$$F_{[L-1,L(p-1)]} \approx p_L{}^*VAR_L/Var_p$$

(where L = the number of litters). Setting $VAR_L = VAR_p$, we obtain

$$F_{[L-1,L(p-1)]} \approx p_L$$

and for values as low as L = 3 and $p_L = 6$, this guarantees an F of 6, which is obtained with a probability of less than 0.05. In other words, randomly sampling 6 pups in each of 3 randomly selected litters should produce significant litter effects, providing only that $VAR_L \approx VAR_p$. This suggests that litter effects are likely to be biologically ubiquitous, particularly given that at least in outbred species $VAR_L \approx VAR_p$ is a rather conservative assumption.

METHOD

Terminology

The reader should think of the experimental situation being simulated as a conventional one-way ANOVA design, with a control and one or more treatment groups. Our model differs from this conventional model only in the introduction of two-stage sampling. That is, the scores or subjects within a treatment group are comprised of measures taken from p pups/litter, and L litters. This design is symbolized in Table 1.

TABLE 1
EXPERIMENTAL DESIGN

Condition	Litter 1	Litter 2	.....	Litter 6	n_c
c_1	$pup_{1,1,1}$	$pup_{1,2,1}$	.....	$pup_{1,6,1}$	
	$pup_{1,1,2}$	$pup_{1,2,2}$	.....	$pup_{1,6,2}$	12
c_2	$pup_{2,1,1}$	$pup_{2,2,1}$	.....	$pup_{2,6,1}$	
	$pup_{2,1,2}$	$pup_{2,2,2}$	.....	$pup_{2,6,2}$	12
c_3	$pup_{3,1,1}$	$pup_{3,2,1}$	.....	$pup_{3,6,1}$	
	$pup_{3,1,2}$	$pup_{3,2,2}$	.....	$pup_{3,6,2}$	12

$N = 36$.
Shown are 3 groups or conditions, each with six litters and 2 pups/litter.

Monte-Carlo Simulation

The simulation program prompted the operator to enter values for S (the number of simulations), C (number of experimental conditions), L_c, p_L, SD_L, SD_p, and the means of the n_c scores in the c groups or conditions. The SD_p and SD_L are held constant over all C cells by the program. Effect size (Z-score) was calculated for a control mean of 100 and an MS_w of 25, with the C–1 treatment means increasing in equal increments to the highest score, which was $(Z^*(MS_w)^{1/2}) + 100$. In other words, for C = 3 and Z = 2.0, the group means were 100, 125, and 150. VAR_L was set equal to VAR_p or in some extremely conservative cases to ¼ VAR_p.

The program generated a random-normal distribution with a mean of 0 and unit variance using RAN 1 (13) which is a portable random number generator with a virtual infinite period and the attribute of lacking sequential correlations. GASDEV (13) controlled the sequence of random numbers and produced a random-normal list of numbers. This routine was used repeatedly to produce the n_c scores per condition. The program first drew a litter mean from a random-normal distribution with mean and SD_L as specified by the operator. The program next drew p_L scores from a random-normal distribution with the chosen litter mean and deviation of SD_p. This procedure was repeated L_c times per cell, for c cells. The operator stipulated whether scores from individual pups were to be averaged to produce 1 score per litter or treated as independent scores, simulating the use of inflated n.

The program next performed a one-way ANOVA on the resulting data set, computed F ratios and the probability of F for the specified degrees of freedom. It then wrote SD and mean for each of the c conditions, MS_w, MS_B, F and p to an output file. This simulation was repeated as often as specified by the operator (1,000 times for all simulations reported herein).

Simulation From Actual Data

An NCTR strain of Sprague–Dawley rats was used to produce 39 untreated litters. Animal care and housing techniques have been reported previously (7). On GD 20 (morning of discovery of copulatory plug = GD 0) all dams were killed by CO_2 suffocation. All fetuses were removed and individually weighed. No litter produced less than 8 fetuses and mean litter size was 12 fetuses. The resulting body weight data set was used for simulations based on actual data.

Again a simulation program was written in C. The operator specified the number of conditions (C), the number of litters per condition (L_c), the number of pups per litter (p_L), the number of simulations (S), and whether or not pup body weights were to be averaged to produce a single score per litter. The option of not averaging within litter simulated the use of inflated n. The program next randomly picked L litters per condition without replacement from the 39 actual litters and then randomly sampled p_L pups from each of the chosen litters, again without replacement. The resulting data set was output and results analyzed as above. This procedure was repeated 1,000 times for a single simulation.

Analysis of Simulation Results

The data sets resulting from either of the aforementioned procedures were analyzed with SAS (14). This allowed computation of within- and between group variance, for comparison to that expected from sampling with a one-stage technique. It

also allowed determination of Type I and II errors, and of obtained power and α levels. In some cases, these were compared to those obtained from one-stage sampling as calculated by a commercially-available program (4). In all cases, comparisons were on percentage of cases which did or did not obtain significance. In this binomial situation, confidence intervals were equal to or less than 3%.

RESULTS

Litter Effects

Litter effects were investigated for L = 3 litters, with p_L ranging from 2 to 12. Monte-Carlo simulations of litter effects were conducted with $SD_L = SD_p = 25$ and an over-all litter mean of 100. Simulations based on the body weight data sampled p_L fetal weights randomly without replacement from 3 randomly-selected litters. (For actual data, p_L did not exceed 8, because one of 39 litters had only 8 fetuses.) Results were virtually identical for the two approaches. Both revealed that litter effects were substantial (Fig. 1). Almost 20% of the time three randomly selected litters differed significantly from each other when as few as two pups were sampled per litter. With as many as 12 pups sampled per litter (and 12 is mean litter size in our laboratory), three randomly-selected litters differed significantly 80% of the time.

The Effect of Inflated n

All simulations were conducted with n_c held constant at 12 scores. L_c varied over the values 1, 2, 3, 4, 6, and 12, whereas p_L correspondingly varied over 12, 6, 4, 3, 2, and 1. Z was set at 0 for the Monte-Carlo simulations, allowing comparison of obtained alpha (the probability of making a Type I or false-positive error) to the nominal $\alpha = 0.05$.

The results were unequivocal. The Monte-Carlo simulations (Fig. 2) showed that when $VAR_L = VAR_p$, even inflating n by using only 2 scores per litter inflated alpha by almost three-fold. In other words, the simulation suggested that under these circumstances, the obtained rate of 14.5 false-positive errors per 100 tests was substantially in excess of the 5 errors set by α. Even reducing VAR_L to ¼ VAR_p (i.e., setting variance between litter means to substantially less than variance within litters) resulted in a Type I error rate of 8.2%, in the limiting case of inflation by 2 scores rather than 1 score per litter.

Again, scores drawn randomly from actual litters and fetuses gave the same results (Fig. 3). Drawing as few as two scores/litter increased the Type I error rate to over 12%; and this rate climbed rapidly, to roughly 50% for 6 fetal weights drawn from each of 2 litters in each of three conditions.

The Effect of Two-Stage Sampling on Power

Monte-Carlo simulations were conducted with the number of conditions c = 3, $L_c = 6$ litters/condition, and $p_L = 1$ pup per litter. Hence, $n_c = 6$, and there were $N = 18$ scores for the entire simulated experiment. Control mean was set at 100, and $SD_L = SD_p = 25$. Z values were calculated for $MS_w = 625$, over a range from 0.5 to 3.0. This meant that means for the maximal effect similarly ranged over 112.5 (Z = 0.5) to 175 (Z = 3.0). For comparison purposes, power (the per-

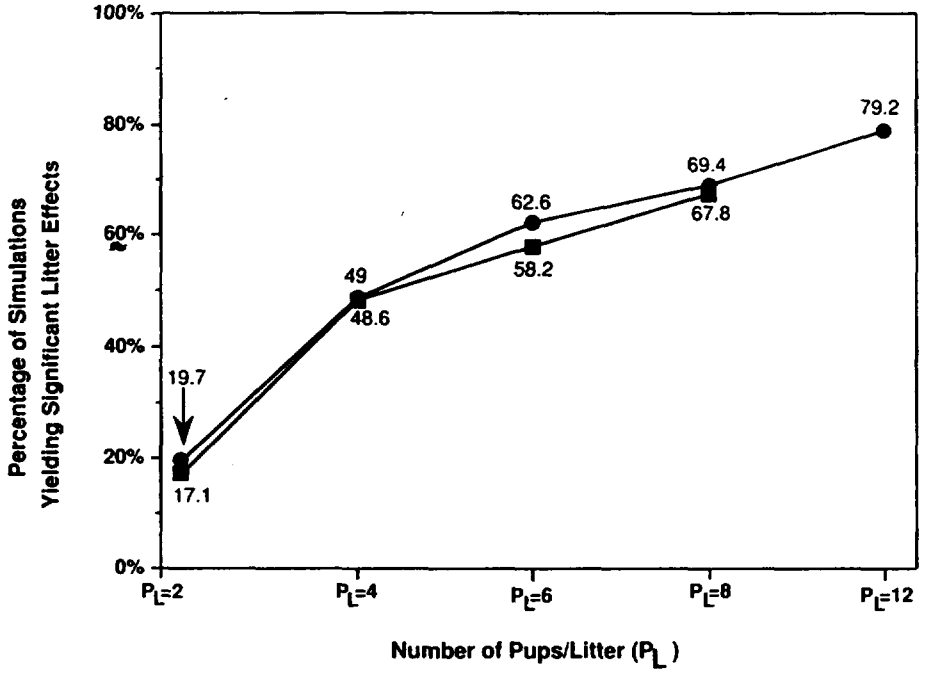

FIG. 1. Litter effects: *Percentage of cases in which 3 randomly selected litters differ significantly*. The percentages shown are for a range of pups per litter (2 to 12 for the Monte-Carlo simulations, 2 to 8 for simulations based on actual data). -●- Monte-Carlo simulations; -■- Simulations from actual data.

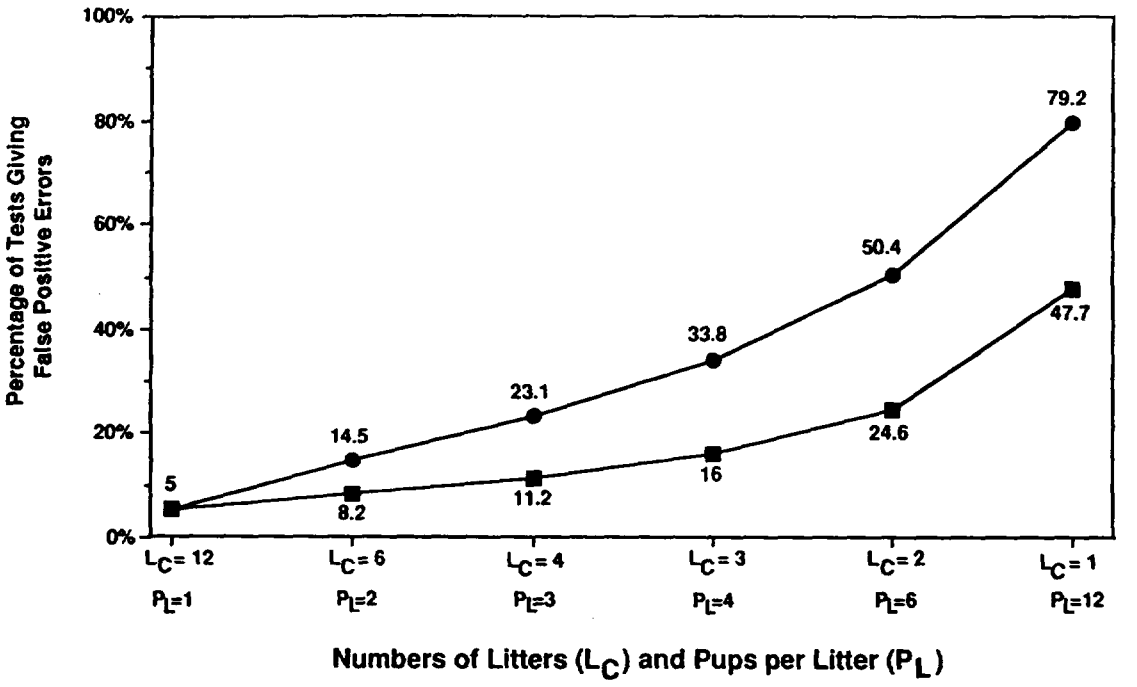

FIG. 2. The effect of inflated n: 1. Monte-Carlo simulations were used to calculate the number of false-positives obtained when n is inflated by 2 to 12 pups/litter. Calculations are conducted for two scenarios; the biologically-probable $VAR_L = VAR_p$, and the extremely conservative $SD_L = \frac{1}{2} SD_p$. -●- $VAR_L = VAR_p$; -■- $VAR_L = \frac{1}{4} VAR_p = \frac{1}{2} SD_p$.

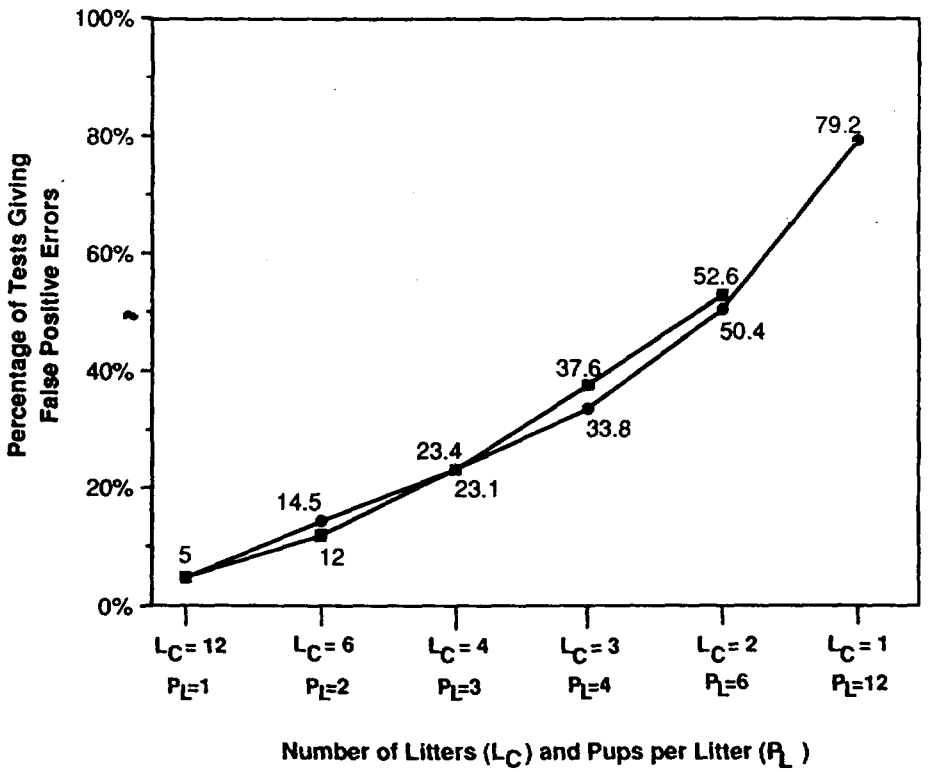

FIG. 3. The effect of inflated n: 2. Values on ordinate and abscissa are as in Fig. 1. Shown are Monte-Carlo simulations of L = 3 litters under the $VAR_L = VAR_p$ conditions, and simulations from actual data sets. -●- Monte-Carlo simulations; -■- Simulations from actual data.

centage of effects correctly identified as significant) was calculated for the above Z scores for single-stage sampling with $MS_w = 625$.

As seen in Fig. 4, two-stage sampling substantially reduces power compared to the one-stage model, even when n is not inflated. This reduction in power can be partially offset by averaging across more than 1 score/litter. This averaging still produces one score/litter, but "smoothes" the data by reducing the impact of getting a single outlier when $p_L = 1/litter$. Indeed, the simulations suggest that for $Z = 1.0$, even averaging scores across 2 pups/litter increases power by some 25%, although using all 12 pups/litter can increase power by 50% (Fig. 5).

However, averaging scores within a litter is not without cost. Measurements per pup are often time consuming, so increasing p_L increases labor, without a corresponding increase in n_c (always providing the experimenter does not inflate n with these additional scores per litter!). It is instructive to compare the results of measuring just one pup per litter to those obtained by averaging more pups per litter from correspondingly fewer litter means. Thus, using 6 litters and 2 scores per litter to produce 6 mean litter scores per condition is compared to using the same number of pups ($2 * 6 = 12$) but with just one pup per litter and $n_c = L_c = 12$ litters. Under these circumstances, we find that one always obtains greater power for the same measurement effort by using just one pup per litter, compared to using (say) ½ as many litters with 2 pups per litter contributing to a single litter mean (Fig. 5).

CONCLUSIONS

One of the most important conclusions derived from this simulation study is that in multiparous species litter effects are very large. These conclusions are based on theoretical considerations, on simulations derived from these considerations, and from simulations based on actual data. They are also strongly supported by other studies which have directly measured litter effects (2,8,9). Thus, as few as three randomly-selected litters may be expected to differ significantly 80% of the time, given that all 12 pups per litter are measured.

The magnitude of litter effects in common laboratory rodents may be attributed to two factors. First, especially in outbred strains, phenotypic variance between litters is naturally high, due to genetic, intrauterine and postnatal maternal care effects. Second, the large litters produced by rodents allow a powerful measurement of within-litter variability, against which between-litter variability can be tested. These two factors combine to assure that within-litter variability is typically not greater than between-litter variability and that powerful tests of these litter effects are available.

Although the magnitude of litter effects found in this study is large, we suspect that, if anything, these estimates are conservative. Typically, the effect of parental genes seems to increase with age, in that animals and humans become more like their parents over time (12). It is not unlikely that in animals older than the GD 20 fetuses used in this study, within-litter variance will become substantially lower than variance between litter means. This effect would, of course, further increase the already sizeable effects of litter.

The predominance of large litter effects has important consequences for experimental design. Perhaps most important, it assures that the practice of inflating n by using more than 1 score/litter will seriously inflate the probability of Type 1 errors. Thus, when litter effects are large, using many pups from a few litters per condition will confound treatment effects with litter effects. As this study has shown, with even the minimal inflation of $n = 2$ scores/litter the probability of Type 1 or

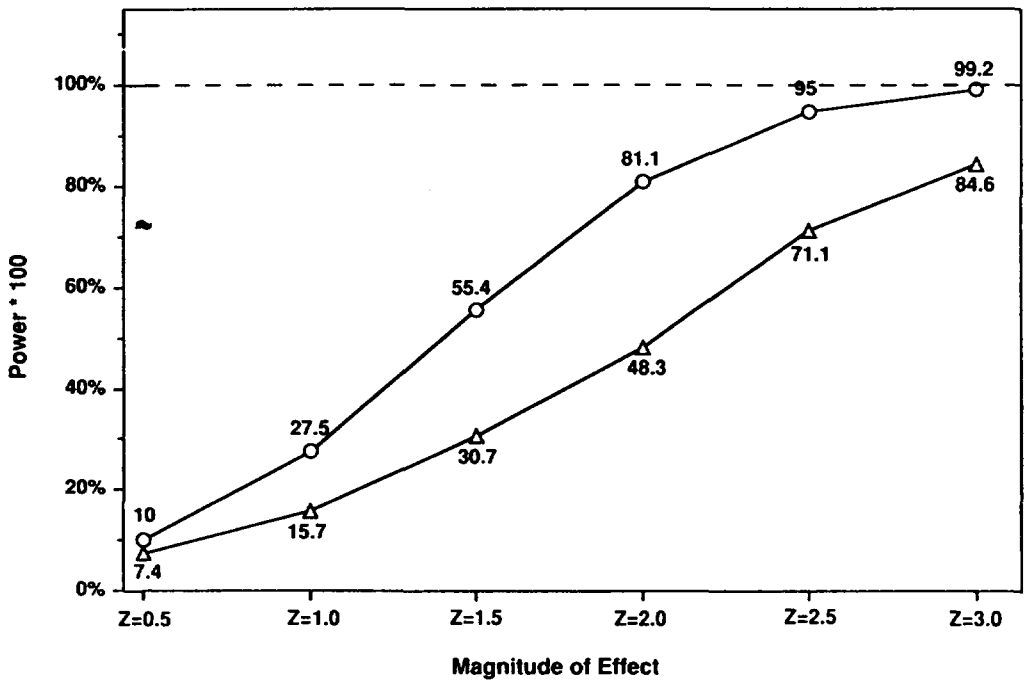

FIG. 4. The effect of two-stage sampling on power. Power (Percent of simulations correctly detecting differences between C = 3 conditions) is shown over a range of effect sizes (Z = 0.5 to 3.0). The effects of one-stage sampling with $MS_w = 625$ are compared to those of two-stage sampling with $SD_L = SD_p = 25$. -O- one stage sampling; -Δ- two-stage sampling.

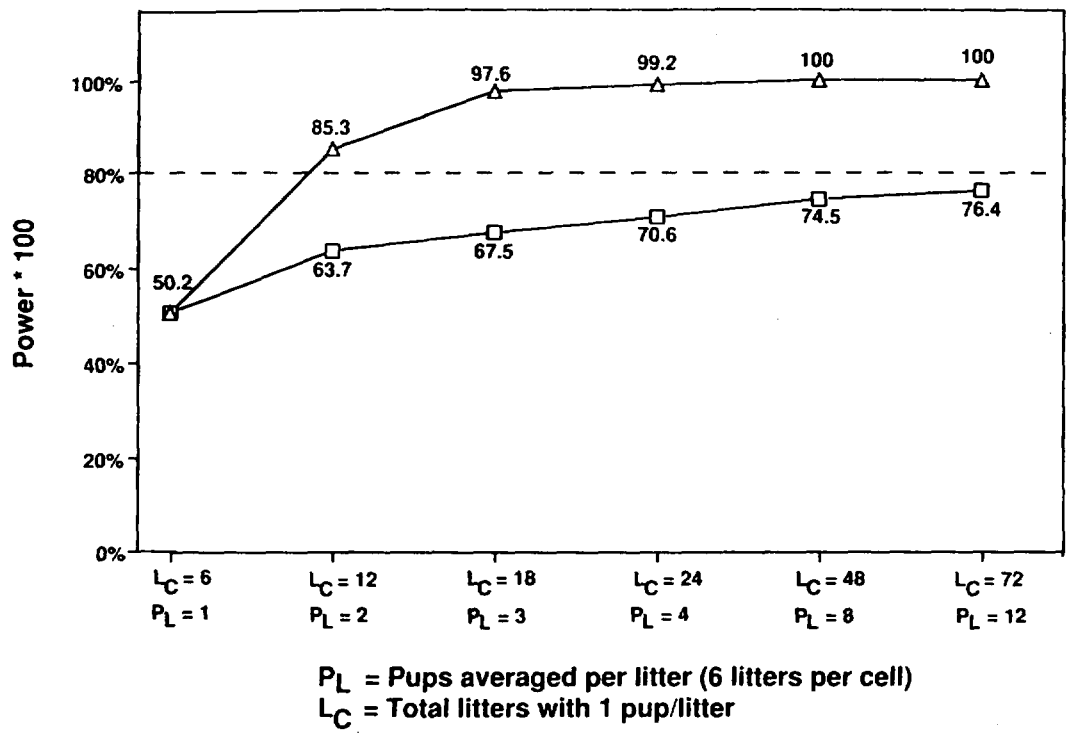

FIG. 5. The effect of averaging within litters on power. Power (Percent of cases in which significant differences were correctly identified) is shown for Z = 2.0, L_c = 6, and p_L = 1 to 12. For the line with the square symbols, all p_L scores are averaged within litter, for n_c = L_c = 6 scores/condition. These results are compared to the effect of holding p_L = 1 and n_c = L_c*p_L; i.e. taking doing the same total number of measures, but from 1 pup per litter.

false-positive errors increased almost three-fold above the traditional 0.05. Still larger litter effects or use of more than 2 pups per litter will further increase the probability of serious error.

Given the potential severity of such violations of assumption, it is important to have some idea of how commonly these errors occur in the neurobehavioral teratology literature. Any truly representative assessment of the frequency of such errors is beyond the scope of this article. However, we did attempt an estimate for one small part of the literature. We assessed all articles published since 1970 (that we have been able to locate!) on the effect of prenatal haloperidol exposure. Table 2 outlines the results. In almost ⅓ of the articles, it was not possible to determine how offspring had been sampled. Of the remainder, over half had inflated *n* by using more than a single score per litter. It is not unlikely, then, that as many as half the publications in this area of research contain such erroneous inflations of sample size, a likelihood that must be viewed with concern.

Turning to the other major result of this study, the availability of more than one offspring per pregnancy also effects the outcome of studies of prenatal exposure effects. It does so by necessitating a two-stage sampling procedure for most measurements; sampling from potential dams, then sampling again over a subset of pups within each litter. The effect of this two-stage procedure is to inflate within-treatment variability (MS_w, the error term) beyond that seen in one-stage sampling. Indeed the expression for variability within a single treatment group with L_c litters and p_L pups/litter is:

$$\frac{p_L (L_c - 1)(VAR_L + VAR_p) + (p_L - 1) VAR_p}{(L_c * p_L) - 1}$$

Clearly, the variance contributed by pups within litters adds substantially to total within-group variance.

This inflation of the error term appeared prominently in our models, and also in data randomly sampled over and within actual litters. Now, one could argue that obtained variance is obtained variance, and the fact that such variance is higher than a theoretical minimum is primarily of theoretical interest. A much more important issue is how, if at all, the experimenter may minimize the error term in this kind of study. We have shown that averaging over more than one pup per litter smoothes the data by reducing the effect of outliers, while maintaining one composite or mean score per litter. The effect of such averaging on power can be quite substantial, because averaging over all pups per litter can increase power by 50%.

Such averages are not without cost, because they generally increase the amount of work. If such factors are limiting, we have also shown that power is always higher when the total number of measurements are aggregated from measures on

TABLE 2

EFFECTS OF PRENATAL HALOPERIDOL

• Inflated *n*	39%
• NO Information on *n*	30.5%
• *n* NOT Inflated	30.5%

Data from 23 papers published since 1970

just one pup per litter. Thus, if it is relatively more difficult to produce litters than it is to measure variables in pups (for body weights, or when tissues can be pooled over a number of pups within a litter), it makes sense to average over many pups per litter. On the other hand, if it is relatively easy to produce treated litters but difficult to collect measures on pups (assessment of operant behaviors in offspring, for example), the optimal research strategy will be to measure not more than one animal per litter.

In conclusion, inflation of sample size by use of more than one score per prenatally treated litter is a serious error. Efforts should be made to disseminate this fact throughout the research community. Such efforts can take a variety of forms, from word of mouth through printed statements of policy to refusal by reviewers to accept for publication articles that contain such errors. Of course, to some degree inflation of sample size is a consequence not just of ignorance but of limited time and resources. It is important for experimenters to understand that averaging within litters is an alternative that can also increase power, if never to the degree obtained by use of the same total number of measurements derived from only one measurement per litter.

ACKNOWLEDGEMENTS

We thank Dr. Ed Riley for suggesting this project. We also would like to thank Mrs. Edwina Martin and Mrs. Loetta Bradford for turning a mess into a manuscript, and Kat Wheeler for producing the figures.

REFERENCES

1. Abbey, H.; Howard, E. Statistical procedures in developmental studies on species with multiple offspring. Develop. Psychobiol. 6:329–335; 1973.
2. Buelke-Sam, J.; Kimmel, C. A.; Adam, J.; Nelson, C. J.; Vorhees, C. V.; Wright, D. C.; St. Omer, V.; Korol, B. A.; Butcher, R. E.; Geyer, M. A.; Holson, J. F.; Kutscher, C. L.; Wayner, M. J. Collaborative behavioral teratology study: Results. Neurotox. Teratol. 7:591–624; 1985.
3. Cochran, W. G. Sampling techniques. New York: Wiley; 1963.
4. Cohen, J. Statistical power analysis for the behavioral sciences. (2nd ed.): Hillsdale, NJ: Erlbaum; 1988.
5. Franklin, J. B.; Brent, R. L. The effect of uterine vascular clamping on the development of rat embryos three to fourteen days old. J. Morphol. 115:273–290; 1964.
6. Henderson, N. D. The confounding effects of genetic variables in early experience: Can we ignore them? Develop. Psychobiol. 1:146–152; 1968.
7. Holson, R.; Adam, J.; Buelke-Sam, J.; Gough, B.; Kimmel, C. d-Amphetamine as a behavioral teratogen: Effects depend on dose, set, age and task. Neurotox. Teratol. 7:753–758; 1985.
8. Hughes, C. W. Outcome of early experience studies as affected by between-litters variance. J. Nutr. 109:642–645; 1979.
9. King, D. L. The effect of early experience and litter on some weight and maturational variables. Develop. Psychobiol. 1:576–584; 1969.
10. Kirk, R. E. Experimental design: Procedures for the behavioral sciences, (2nd ed.). Belmont, CA: Brooks/Cole; 1982.
11. Lindman, H. R. Analysis of variance in complex experimental designs. San Francisco:Freeman; 1974:35–36.
12. Nonaka, K.; Matsumoto, T.; Nakata, M. Genetic and environmental factors in the longitudinal growth of rat: 1. Body weight and overall craniofacial size. J. Craniofacial Genet. Develop. Biol. 8:319–327; 1988.
13. Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. Numerical recipes in C: The art of scientific computing. Cambridge, Cambridge University Press; 1986.
14. SAS Institute Inc. SAS/STAT, Guide for Personal Computers, Version 6. Cary, NC: 1987.
15. Winer, B. J. Statistical principles in experimental design. New York: McGraw-Hill; 1971.